RESEARCH ON THE TECHNOLOGY OF INFERENCE AND DECISION

Ward Edwards

Michigan University

Prepared for:

Office of Naval Research
Advanced Research Projects Agency

30 November 1973

011313-F

THE UNIVERSITY OF MICHIGAN

ENGINEERING PSYCHOLOGY LABORATORY

Final Technical Report

Research on the Technology of Inference and Decision

WARD EDWARDS
Principal Investigator

Sponsored by:
Advanced Research Projects Agency
ARPA ORDER No. 2105

Monitored by:
Engineering Psychology Programs
Office of Naval Research
Contract No. N00014-67-A-0181-0049
NR 197-021

Contract Period: 1 June 1972 - 31 May 1974

Administered through:

November 1973

OFFICE OF RESEARCH ADMINISTRATION · ANN ARBOR

011313-F

RESEARCH ON THE TECHNOLOGY OF INFERENCE AND DECISION

Final Technical Report

30 November 1973

Ward Edwards

Engineering Psychology Laboratory
The University of Michigan
Ann Arbor, Michigan

This is a final report of research supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Engineering Psychology Programs, Office of Naval Research under Contract No. N00014-67-A-0181-0049, Work Unit Number No. 197-021.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 011313-F	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) RESEARCH ON THE TECHNOLOGY OF INFERENCE AND DECISION		5. TYPE OF REPORT & PERIOD COVERED Final Technical 1 June 1972 - 30 November 1973 6. PERFORMING ORG. REPORT NUMBER None
7. AUTHOR(s) Ward Edwards		8. CONTRACT OR GRANT NUMBER(s) N00014-67-A-0181-0049
9. PERFORMING ORGANIZATION NAME AND ADDRESS Engineering Psychology Laboratory Institute of Science & Technology University of Michigan Ann Arbor, Michigan 48105		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 197-021 ARPA Order No. 2105
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia		12. REPORT DATE 30 November 1973 13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME AND ADDRESS (if different from Controlling Office) Engineering Psychology Programs Office of Naval Research Department of the Navy Arlington, Virginia		15. SECURITY CLASS. (of this report) Unclassified 15a DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Flat maxima
Multiattribute utility theory
Elicitation
Information processing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report summarizes sixteen months of research on the technology of inference and decision. Efforts in arranging a conference involving the national security community and research personnel in the summer of 1973 are detailed. The first major research result, flat maxima in decision analysis, is summarized and its implications for social and psychological theory as well as decision technology are discussed. The major conclusions from this area of research is that the structuring of the decision problem may be much more important than the elicitation

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

of specific parameter values. The second major topic is a review of multi-attribute utility measurement and an experiment is outlined which may circumvent the problem of evaluating utility elicitation methodologies without having an external standard for utility. Elicitation technology for probabilities constitutes the third research area. Extensive review of data collected in five critical experiments indicated that feedback may be more important than aggregation as a source of human conservatism in inference. The importance of this finding is stressed. The fourth and last area summarizes the final three of nine Technical Reports to stem from this contract. These concern Bayesian statistical analysis for comparing deterministic models, a critique of the Bayesian foundations of decision theory, and error analysis in Probabilistic Information Processing systems.

TABLE OF CONTENTS

INTRODUCTION.....1

A TECHNICAL OVERVIEW

Application of decision technology to actual military decisions.....2

Flat maxima in decision analysis.....5

Multiattribute utility measurement.....11

Elicitation technology for probabilities.....16

Assorted criticisms.....20

MANAGEMENT INFORMATION.....21

REFERENCES.....24

ABSTRACTS

Introduction

On 18 January 1972 the University of Michigan submitted to the Advanced Research Projects Agency a proposal for Research on the Technology of Inference and Decision. The proposal called for 5 years of research at a total cost of $400,000. The Principal Investigator was Professor Ward Edwards. A one-year contract was awarded, N00014-67-0181-0049, and research began on 10 June 1972, the date on which funds became available. Monitoring responsibility for the contract was undertaken by Dr. Martin A. Tolcott of the Office of Naval Research. In January of 1973 a proposal for continuation of the program was submitted to ARPA, and was funded. In March of 1973 it began to seem likely that Edwards would leave the University of Michigan, and in May he accepted an offer to become Director of the Social Science Research Institute, University of Southern California. Plans were then made to terminate the contract with Michigan. Consequently, although this is formally a final report of the work at Michigan, it is in fact a progress report covering about 1 1/3 years of a five-year program.

The main products of the program so far have been nine technical reports which are receiving distribution independently of this Final Report. They stand alone and speak for themselves. Consequently the purpose of this Final Report will be to present the overall conception into which the technical reports fit, to report on incomplete activities that will continue at USC, and to summarize

some financial and personnel information. Abstracts of the nine technical reports are also included.

A Technical Overveiw

The original proposal that led to this program called for research bearing on the topics of information processing (especially in intelligence systems), of tactical action selection, and of information acquisition. Research was to be a mixture of theoretical work, laboratory work, and work growing out of contacts with real military environments. As the program developed, four main themes emerged, and in addition some other activities less closely related to these also occured. So this review will be organized under five headings, of which the last is somewhat of a catch-all.

Application of decision technology to actual military decisions. In late 1972 and early 1973, Col. Kibler, of ARPA, and Edwards had several conversations about how ARPA should go about encouraging the development of decision technology in paths relevant to military problems and its application at relatively high levels within the national security community. Evidence of applicability, obtained mostly within the intelligence community, exists in reasonable abundance. But the nature of the technology is such as to produce relatively high levels of resistance to application, and so evidence of successful application within one agency is not enough to produce

efforts at application within other agencies (or even elsewhere within the same agency).

The upshot of these discussions was a decision to hold a summer study in the summer of 1973, under ARPA sponsorship. Attendees would be high-level members of the national security community, who would brief the study group about the kinds of decision problems their organizations encounter and the methods they use to solve them, and key individuals within the academic and scientific community, who would brief the study group about technological tools available or in prospect, and about research needs. The goal would be the production of a document that would suggest appropriate directions for subsequent ARPA activities in the field. Col. Kibler asked Dr. Davis Bobrow, of the University of Minnesota, and Edwards to take joint responsibility for the scientific leadership of the study. Edwards's activities prior to the conference were sponsored by this contract, although the conference itself (and Edwards's activities while attending it) were supported independently.

This turned out to be a substantial job. It required Edwards to travel to Washington for conferences with Kibler and Bobrow and others on February 9, January 10, March 20 and May 1, to travel to Minneapolis on June 14, and to receive a number of visits in Ann Arbor. Interactions with invitees took time. Most important,

however, was the writing, jointly with Bobrow, of a "straw man" study report. This document was distributed to the participants prior to or at the beginning of the study as a guide to what Bobrow and Edwards considered important and appropriate. While the final study report was quite different from the "straw man" version, it seems safe to say that a good deal of the agenda of the study reflected various kinds of reactions to the straw man, and consequently that that document played a significant role in engendering the final output of the study.

It is not appropriate to distribute the "straw man" report as a technical report of this contract, since it was not intended for such general distribution. Nor is it appropriate to include it in this final report, both because it is too long and because it was not intended to have that kind of performance. Nevertheless, in terms of its influence it seems possible that it is the most significant output of this first 1 1/3 years of this program.

The summer study itself in various ways furthered the objectives of this contract. It established relations between Edwards and a number of members of the national security community whom he had not previously known, many in a position to provide access to decision settings and systems of considerable national importance. In particular, it highlighted the nature and importance of WWMCCS, and permitted Edwards to become acquainted with at least a few of

those involved. On the basis of what was learned at that study, WWNCCS seems like a sufficiently promising locus for the application of decision technology to deserve a much closer look; such a look should be taken early in 1974.

Flat maxima in decision analysis. The phenomenon of flat maxima has been noticed by decision analysts looking at a number of specific contexts, ranging from the use of proper scoring rules in weather forecasting (Murphy and Winkler, 1970), to information purchase with optional stopping (Edwards, 1965). While everyone has assumed that these were special instances of a much more general phenomenon, no one has attempted to define the general phenomenon or to look at the broad range of its implications.

As a first step in looking into the question, von Winterfeldt and Edwards looked at the applications of decision-theoretical thinking to sensory processes and to probability estimation. The result was the first Technical Report of this program; its abstract appears in the section at the end of this final report that contains the abstracts of all technical reports so far produced. (They are listed there in the order in which they are discussed here.)

The method used by von Winterfeldt and Edwards in that study was mathematical, but rather specific. While it did not work with only specific examples, it proved no general theorems. Nevertheless,

it was able to reach some general conclusions. The basic conclusion is that everywhere one looks, decision-theoretical maxima are flat; that is, significant deviations from optimal strategy lead to relatively insignificant percentage reductions in expected payoff.

After drafting that report, von Winterfeldt and Edwards continued to gnaw away at the problem. How can the flat-maximum problem be formulated generally, rather than as a set of specific examples? The solution had to lie in the convexity property of decision-theoretical payoff functions--and it did. Ultimately, von Winterfeldt and Edwards were able to formulate and prove the general theorem which the examples all exemplify. (See Technical Report Abstract No. 2.)

What does the flatness property mean? It is easy to overinterpret it. The difference in expected payoff between an optimal strategy and a non-adjacent suboptimal one can be made as large as desired by simply increasing the magnitudes of all payoffs. If the expected payoff associated with an optimal strategy is a billion dollars, then a 1% reduction in that expected payoff is ten million dollars. Flatness is meaningful only when considered in percentage terms, as the first von Winterfeldt-Edwards report makes clear.

The implications of flatness can be looked at in two ways: from the point of view of the decision analyst, or from the point

of view of the social philosopher. Consider the latter first. From this point of view, the relative insensitivity of payoffs to significant but not monstrous deviations from optimal behavior is a kind of glue that permits society to hold together. Suppose that the consequences of even minor deviations from complete rationality were grossly disastrous--how long could a society of only-partly-rational men survive? But the fact that minor deviations are almost costless leaves some room for both error and individual differences, while the fact that they are not completely costless makes analysis and intelligence worth bringing to bear on decisions.

The decision analyst must face much more specific consequences of flatness. These consequences fall most heavily on elicitation technology. For example, the use of proper scoring rules has been assumed to motivate probability estimators to produce "good" estimates. But analysis shows that relatively large deviations from the optimal probability estimate produce only relatively small reductions in expected payoff. Consequently, the motivating effect of proper scoring rules reviewed below shows exactly that. Experiments have shown that proper scoring rules improve probability estimates--but they certianly have not established that it is the rules themselves, rather than the indoctrination and practice that goes with their use, that cause the improvement.

On the other hand, the fact of flat maxima makes precise probability and value estimates of less importance than they might otherwise be. If a 10% error in an estimated quantity produces

only a 2% decrease in expected value, perhaps that 10% error is tolerable--certainly more tolerable than if it produced a 30% decrease in expected value.

The implication of this argument is that the most important aspect of decision analysis is the structuring of the problem for analysis, not the elicitation of numbers and computational processes that follow. Unfortunately, this process of structuring the problem is least amenable to formal prescription. It seems to be mostly a matter of wisdom, experience, and ability to tolerate confusion, ambiguity, and conflict.

For psychological theory, the phenomenon of flat maxima has yet another implication. Psychological models, such as probabilistic learning models, incorporate two classes of parameters: known parameters, such as the number of stimuli used in an experiment, and parameters to be estimated from data, such as learning rates. Known parameters are errorless. But estimated parameters are always estimated via decision-theoretical procedures such as maximum likelihood, least squares, etc. These procedures formally have the properties of all decision-theoretical flat maxima. Consequently, the appearance of precision given by, say, a least squares estimate of a parameter is somewhat spurious--substantially different values of the parameter would produce only modest increases in the sum of squares that was used as criterion for parameter estimation.

In an attempt to evalute the effect of this decision-theoretical phenomenon on learning models, von Winterfeldt and Edwards have compared errors produced by inaccurate parameter estimates with errors produc by erroneous values of known stimulus parameters. Very generally, the models are far more sensitive to the latter kinds of errors than to the former. To put if another way, learning models depend very sensitively on numbers that describe the environment, but only very weakly on numbers that describe the organism.

To what extent, then, are they models of organisms? The traditional distinction between normative and descriptive models is that normative models describe tasks, while descriptive models describe what organisms do in tasks. But if the description of what organisms do in tasks is vague, in the sense that a different description produces almost the same result, then why should theorists accept relatively poor descriptions of tasks instead of simply using the appropriate normative models?

The argument sketched above is not yet fully formed. From here on, the issues get more philosophical and less technological. They have to do with model success. What is a mode. for? How can you tell when it is doing what it should do? A substantial philosophical literature exists on this question, and we have explored it, but found it unhelpful. In our mathematical study

of the parameter estimation process, however, we find the curious result that while least squares, chi square, and similar estimation procedures have the flatness that we expect, maximum likelihood procedures do not. In fact, estimates based on maximum likelihood procedures seem to be unreasonably sharp--implying more precision than the data could ever in fact yield. We had hoped to have a technical report ready on this set of topics by the end of this contract, and in fact a draft version exists. But these issues remain sufficiently unclear to us that we have decided not to issue it. In Los Angeles, we hope to get some advice from R. Duncan Luce about these difficult questions.

We have pursued one further line of thought in this area. Some nondecision-theoretical transformation clearly can restore sharpness to flat maxima. An obvious example might be called the winner transformation. While the loss function associated with deviation from optimal probability estimates when a proper scoring rule is being used is inevitably flat, sharpness can be restored by turning the situation into a contest. For example, weather forecasters might compete for a week, and whichever one had the highest cumulative Brier score at the end of the week might win a prize.

The trouble with the winner transformation, unfortunately, is that its decision-theoretical properties are unpleasant. It is

an instance of a widespread phenomenon of real life--the implicit linear scoring rule. In a single-estimate contest, for example, the optimal strategy under the winner transformation is to estimate the most likely event as having probability 1, and all its competitors as having probability 0 (unless the estimator can know his competitors' estimates when he makes his own). Strategies get much more complicated in a multiple-estimate contests, but they all have this sort of flavor.

Still, these unpleasant formal properties may not be sufficiently good reason to reject the winner transformation as a practical tool. Wendt planned such an experiment, in a Bayesian revision task of the two-normal-distributions type. However, as a result of the decision to move to USC, this experiment has been postponed until after the move. This experiment will probably be a part of a larger experiment on response modes and training techniques.

Multiattribute Utility Measurement. The literature and the technology of multiattribute utility measurement have grown very rapidly in the last few years. (See for example Raiffa, 1968, 1969; Keeney, 1971; Edwards, 1971.) Our own view is that this technology stands now roughly where the Bayesian technology stood in 1963--but has a brighter future, because the topic of values and evaluation is inherently more important than the topic of diagnostic inference.

The technique for multiattribute utility measurement, in its rating-scale version, looks deceptively simple (Edwards, 1971). But in fact some rather sophisticated mathematics and some rather strong assumptions lie behind it. In particular, the distinction between riskless and risky multiattribute utility measures, almost meaningless in practical applications (because all practical situations involve some risk, yet it is often not worth while to take it explicitly into account), is strong and important in the underlying theory.

Von Winterfeldt and Fischer (see Technical Report Abstract No. 3) have reviewed the literature bearing on the assumptions that underlie multiattribute utility measurement and the relation of those assumptions to the choice of an elicitation technique. A point that emerges from the review, not so much as a conclusion but as a fact of life, is that the elicitation techniques that are formally justified by the assumptions are far clumsier and more unpleasant to use than one would wish, while the simplest elicitation techniques require very strong assumptions, and even then are less strongly related to the model than might be desired.

The obvious consequence of this state of affairs is that a more serious study should be made than has been of the degree to which multiattribute utility measures based on simple elicitation

techniques agree with those based on more complex techniques. That is, instead of treating utilities as formal numbers, either exactly correct or else useless, one should think of them as approximations, and explore how good various approximations are. Numerical explorations of this question are clearly called for, but will not occur until after the move to USC.

Von Winterfeldt and Edwards (see Technical Report Abstract No. 4) performed an experimental study of several approaches to multiattribute utility measurement. The main finding was that elicitation methods based on gambles were preferable to other elicitation methods. This is a surprising conclusion, since the whole thrust of the choices-among-bets literature is that such choices are poorly linked to the input parameters. It will need further examination.

The most urgent task in the study of utility measurement is the development of what, in our laboratory slang, we call God's utility function--that is, an objective standard with which to compare elicited utiliites. Most of the most important conclusions in the area of probability elicitation have emerged from comparison of elicited probabilities with calculated ones, in situations in which such calculations are possible. In the absence of God's utility function, such comparisons are not possible for utilities--and research is severely handicapped.

An extreme subjectivist would assert that one cannot dispute, or prescribe, tastes--that the goal of finding a situation in which such an external standard can be defined is unattainable. (He might, in fact, make exactly the same argument about probabilities--why are opinions any more prescribable than tastes?) We think that utilities are contextual, and indeed are often interpersonal, and thus are sometimes subject to prescription.

Several approaches to the problem can be conceived of. We have explored one in considerable detail. Diamonds are extremely interesting stimuli for use in utility experiments. They are valuable, and the value is rather precisely reflected in the price, which can be taken as an index of overall utility. The dimensions of value are extremely well specified and understood. They are cut, clarity, color and carats. Of these, all (except perhaps clarity) are in principle objectively measureable--but in practice an appraiser of diamonds works with only a scale, a pair of locking tweezers, a jeweler's loupe, a good, uniform light source, and his highly educated eyes and brain. These experts are extremely highly practiced; a typical wholesale diamond merchant will appraise many thousands of diamonds in the course of a year. We are under the impression that the results of such appraisal show a great deal of inter-expert agreement, though the evidence on the point if less abundant than we might like (see Bruton, 1970).

We plan, and have conducted extensive preliminary work looking toward, an experiment on elicitation technology for multi-attribute utility, using diamond wholesalers as our experts and wholistic judgments of price as the standard of God's utility. The argument will be that the technique that most nearly reproduces (up to a linear transformation) those prices is the best technique. We have arranged for cooperation from a group of diamond wholesalers in New York. (We may be able to obtain judgments in Los Angeles also.) Wendt, who is collaborating with Edwards on this, has returned to Hamburg, and believes he may be able to obtain judgments from diamond wholesalers in Amsterdam, the center of the world diamond market.

The diamond study is the best approach we have yet found to God's utility function. But the stimuli are hard to obtain, and access to the experts is a problem. So we are also exploring the possibility of using the additive nature of certain kinds of objects as the basis for such experiments. A market basket full of groceries is, in a sense, a commodity. But its utility would be conceded by most to be an additive aggregate of the utilities of the objects in the basket. Given the utilities of these separate objects, their sum specifies a form of God's utility function for such baskets. This can be compared with judgmental utilities obtained in one way or another. There are some serious difficulties with this idea,

and it will take considerable further thinking to refine it to the point at which an experiment can grow out of it.

Elicitation technology for probabilities. The topic of elicitation of probabilities has been a major theme of this laboratory's work for more than ten years--and is still by no means a closed topic. The alternatives are pretty well understood, but by no means is there enough information about them, especially with populations other than college students and stimuli other than the typical bookbags-and-poker-chips or pick-up sticks, to permit unhesitating choice among them.

Goodman (see Technical Report Abstract No. 5) has performed very extensive further analyses of the data from the five key experiments done in this laboratory on the topic. Her statistical technique was a form of regression analysis that has not been used in this context before, and the independent variables she studied were: the response mode itself, the scale (log or linear) on which it was expressed, whether or not the subject had to aggregate evidence in his head, and whether or not feedback concerning the meaning of the response was present at the time of response. Her main conclusions are summarized in the abstract. The importance of feedback in producing conservatism had been suspected before--but this analysis is surprising in indicating that that is more

important than whether or not the subject aggregates evidence in his head. The significance of this study of Goodman's to the designers of probabilistic information processing systems would be difficult to overestimate.

Goodman's paper invites an editorial comment. It is reasonably short--and quite difficult to read, mostly because it is full of jargon and very tightly reasoned arguments. Only someone thoroughly steeped in the Bayesian point of view and the Bayesian literature will find it easy to get through. And the cost of writing it in such a way as to make it intelligible to those not already familiar with Bayesian ideas is prohibitive--the length of the paper would triple or quadruple, and much of it would then consist of reviewing familiar ideas. This, of course, is the normal course of development of a field of science--but this report dramatizes the fact that the Bayesian point of view has moved a long way in ten years.

The standard decision-analytic technology of elicitation emphasizes internal consistency. If a subject makes inconsistent judgments, the inconsistency is called to his attention and he is invited to revise any or all judgments to eliminate it (see Raiffa, 1968).

This would be fine if only consistency were important. But, especially when the judgments concern probabilities, veridicality is more so. People tend to be much more secure and confident about posterior odds judgments than about likelihood ratio judgments (Edwards, Phillips, Hays, and Goodman, 1968)--yet the evidence is abundant that likelihood ratio estimates are usually much more accurate than are posterior odds (see Edwards, 1968). Consequently, if a subject judges both, he is likely to be inconsistent (i.e. violate Bayes's theorem). If he is then invited to revise for consistency, it seems quite possible that he will revise, not the odds, but the likelihood ratios, and therefore revise them away from veridicality.

An experiment was designed to explore this hypothesis. The standard pick-up-stick task was used. Subjects first estimated single-stick likelihood ratios, and then estimated posterior odds for four-stick sequences. Then they were taught about Bayes's theorem, their inconsistencies were exhibited, and they were invited to make whatever revisions seemed appropriate.

The data from 15 subjects were highly unsatisfactory. Half of the subjects were more veridical after revision than before; half were less. The problem, we now suspect, is that four-stick sequences are too short to produce sufficiently conservative posterior odds. We plan to start this experiment all over again at

USC, using longer sequences and perhaps a larger value of d'.

Seghers (see Technical Report Abstract No. 6) has conducted an experiment on proper scoring rules taken as bets. Do subjects maximized expected monetary value in such situations? The experiment fairly conclusively says no. This finding means that the assumption that subjects will maximize expected monetary value, usually taken as the basis for use of proper scoring rules, is simply not appropriate. Proper scoring rules probably help to instruct subjects about the meaning of probability estimates, but by themselves they do not constrain the subjects to produce such estimates as their formal nature would prescribe. (The facts about flat maxima make this conclusion all the more reasonable.)

We have been thinking about training, though the thinking is as yet too unstructured to lead to experiments or theory. The first and most important point is that training for probability estimators might be called The Illusory Panacea. Whenever some peculiarity of human behavior in probability estimation situations is noticed, the explanation always is that the estimators were untrained or undertrained. But what constitutes correct and sufficient training? When professional probabilists and Bayesians so often find themselves caught in logical errors, can any lesser standard of training be called enough?

We are coming to think that all this emphasis on training is misplaced. Instead, what now seems important is the packaging of decision technology--that is, specification of a combination of analytic techniques, elicitation methods, and training for judges that will permit its application in practical situations. In short, rather than training everyone to be Julia Childs we are coming to feel that it would be better to have a cookbook--with pretested recipes, please.

Assorted criticisms. Three other Technical Reports are in various ways auxiliary to or critical of work reviewed above. Perhaps the most useful is a report by Wendt (see Technical Report Abstract No. 7) that provides practical techniques for applying the techniques of Bayesian statistical analysis to comparisons among models--especially among deterministic models. This paper grew out of thought about data analysis problems that arose in Seghers' experiment, and is applied in that Report, but the techniques presented by Wendt are far more widely useful than that.

Wendt (see Technical Report Abstract No. 8) is critical of many of the assumptions and working hypotheses of the Bayesian point of view, and has attempted to assemble his criticisms in a coherent

form. This Report is frankly a think-piece. Few decision theorists would agree with all of its arguments: some would disagree with all of them . But foundations should not be allowed to remain unexamined, and criticisms of this sort often lead to later constructive work.

Fryback and Edwards (see Technical Report Abstract No. 9) have looked into the problem of errors produced by variability of likelihood ratio estimates in a probabilistic information processing system. That error can in principle be of substantial size, as a model based on the assumption that such errors are normally distributed makes clear. But actual examination of test-retest reliabilities in a major earlier experiment (Edwards, Phillips, Hays, and Goodman, 1968) makes clear that in fact such reliabilities are very high indeed, and consequently that the effect (at least in that experiment) was relatively small.

Management Information

The following Table lists those who have worked on this contract for significant fractions of time.

Personnel (Based on 16-month period)

Name	Title	MM	
Ward Edwards	Principal Investigator	5.1	
Barbara C. Goodman	Associate Research Psychologist	8.0	
Dirk Wendt	Visiting Research Psychologist	4.8	
Gregoary W. Fischer	Assistant Research Psychologist	7.2	
Kurt Snapper	Research Assistant	.15	
Detlof von Winterfeldt	Research Assistant	15.0	
Dennis G. Fryback	Research Assistant	.25	} no charge to ARPA
Raymond C. Seghers	Research Assistant	.15	
Patricia Homan	Secretary	8.0	
Annette Johnson	Administrative Assistant	5.8	

The following presents financial facts about the work so far on the Contract.

Balance as of 1 October 1973: $63,109.25

Anticipated Expenditures:

October	
Salaries & Wages: (EPL)	$ 3,547.00
Tech. typists	480.00
Graphics	95.00
	4,122.00
I.C. (58.2%)	2,399.00
Total	$ 6,521.00
Direct Costs:	
Supplies:	32.10
Telephone:	25.00
	$ 6,578.10
November	
Salaries & Wages	452.00
Graphics	14.00
Reproduction	308.00
	774.00
I.C. (58.2%)	450.00
Total	1,224.00
Direct Costs:	
Reproduction supplies	225.00
Telephone	5.00
Postage	155.00
	$ 1,609.00

Balance:	$ 63,109.25
Anticipated expenditures:	8,187.10
Estimated balance:	$ 54,922.15

References

Bruton, E. Diamonds. N. A. G. Press Ltd., London, 1970.

Edwards, W. Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. J. math. Psychol., 1965, 2, 312-319.

Edwards, W. Conservatism in human information processing. In Kleinmuntz, B. (Ed.) Formal representation of human judgment. New York:Wiley, 1968.

Edwards, W. Social utilities. In Decision and Risk Analysis: Powerful new tools for management. Proceedings of the sixth triennial Symposium, June, 1971.

Edwards, W., Phillips, L.D., Hays, W.L., & Goodman, B.C. Probabilistic information processing systems: Design and evaluation. IEEE Trans. Syst. Sci. Cybernetics, 1968, 248-265.

Keeney, R. Utility independence and preference for multi-attributed consequences. Operations Research, 1971, 19.

Murphy, A.H., and Winkler, R.L. Scoring rules in probability assessment evaluation. Acta Psychologica, 1970, 34, 273-286.

Raiffa, H. Decision Analysis. Boston:Addison Wesley, 1968.

Raiffa, H. Preferences for multi-attributed alternatives. RAND Memorandum No. 5868. The RAND Corporation, Santa Monica, California, 1969.

011313-1-T

Report Abstract 1

Costs and Payoffs in Perceptual Research

Detlof v. Winterfeldt and Ward Edwards

Abstract

A persistent problem in any kind of psychological research that reaches conclusions about inaccessible processes or experiences inside a subject's head is to validate those conclusions--that is, to exhibit persuasive reasons to believe that emitted behavior is some sense faithfully reports inaccessible processes. In the mid-1950s, perceptual researchers widely adopted an approach that might be called validation by cupidity. If the experimenter is willing to define a correct response, he can reward the subject for correct responses and not for wrong ones; suitable reward schemes combine with an assumption of rational behavior on tne subject's part to permit direct inference of internal processes. However, decision-theoretical maxima are flat, in the sense that seriously inappropriate behavior produces relatively little reduction in the subject's expected payoff. This means that costs and payoffs are rather feeble means of instructing subjects what to do, or of ensuring that he does it.

This argument is made specific in examples drawn from three kinds of perceptual experiments. In some tasks, such as probability estimation, subjects directly estimate subjective quantities, and receive rewards for accuracy of estimate. An analysis of proper scoring rules for probability estimation shows that their maxima are inevitably quite flat. An analysis of a yes-no decision task shows that the incorrect answer produces flat maxima; while the payoff function can be sharpened by increasing the magnitudes of all payoffs, a suitable relative payoff function is intractable. In such yes-no tasks, criterion variability produces even more flatness, so much so that it would be surprising if such variation did not

occur in most real experiments. Criterion variability sufficient to produce a 30% reduction in estimates of d' produce only 5% to 8% reductions in expected winnings.

Implications of these results for experimental design, for interpreting experimental results, and for more general decision-theoretical thinking are discussed.

011313-4-T

Report Abstract 2

FLAT MAXIMA IN LINEAR OPTIMIZATION MODELS

Detlof v. Winterfeldt and Ward Edwards

Abstract

Expected value functions as functions of decisions and decision strategies are flat around their maxima. This so called flat maximum phenomenon has been discovered in sensitivity analyses in virtually all decision theoretic paradigms. But until now most of the research on flat maxima explored more or less general examples and limiting considerations. Two basic questions remained unanswered: what are the mathematical reasons for the restricted shape of the evaluation functions; and can these restrictions be interpreted as flatness in a psychological sense? While the second question calls for psychological experimentation, the first question can be answered with mathematical tools. The present article shows that the mathematical characteristics of linear optimization models impose severe restrictions on the functions evaluating choice alternatives such as gambles, multi-attributed outcomes, or consumption streams. The course of proof of this argument provides a helpful tool for sensitivity analyses in decision theory. The concepts and methods are demonstrated in examples from statistical decision theory, psychological modeling, and applied decision theory.

011313-7-T

Report Abstract 3

MULTI-ATTRIBUTE UTILITY THEORY : MODELS AND ASSESSMENT PROCEDURES

Detlof v. Winterfeldt and Gregory W. Fischer

University of Michigan

Abstract

This article reviews multi-attribute utility theory from a measurement theoretic perspective. It describes and classifies decision situations according to three salient aspects of choice : uncertainty, time-variability, and multi-dimensionality. For each choice situation the main mathematical representations, their interrelations and differences are discussed. Measurement theoretic tests are described which separate between multi-attribute utility models in riskless and risky time invariant choice situations. Assessment procedures are outlined to encode utility functions for the representations developed, and experimental applications of multi-attribute utility theory are briefly reviewed.

Report Abstract 4

Evaluation of Complex Stimuli Using Multi-attribute Utility Procedures

Detlof v. Winterfeldt and Ward Edwards

Three procedures for constructing additive multi-attribute utility theory (MAUT) models were tested for their differential validity: a probabilistic procedure, a simple direct rating procedure, and a modified direct rating procedure. Validation criteria were ratings and simple choices. Procedures were evaluated after an intensive training session in which subjects learned to adopt an evaluation strategy with which they felt most comfortable and which best reflected their preferences. The results of a correlational analysis indicated that MAUT can improve upon the decision maker's own unaided intuition. The probabilistic procedure was found to be the superior method for predicting simple choices between stimuli.

011313-5-T

Report Abstract 5

Direct Estimation Procedures for Eliciting Judgments about Uncertain Events

Barbara C. Goodman

This report re-analyses data from five studies concerned with methods for eliciting judgments about uncertain events. It focusses on response modes, such as odds, likelihood ratios, etc.; whether or not the response required the subject to aggregate items of data in his head; whether the scale on which the response was made logarithmic or linear; and whether the subject received additional feedback about the implications of his estimates in the course of making them. While no single experiment studies all these issues simultaneously, combination of the data from the five experiments permits some strong conclusions:

1. Presence of additional feedback about the implications of estimates is probably the most powerful variable controlling the extremeness of these estimates; feedback makes estimates less extreme. Whether the less extreme estimates are closer to or further from correct Bayesian values depends on stimulus conditions.

2. Aggregated responses are consistently less extreme than nonaggregated responses.

3. Linear scales produce less extreme responses than logarithmic scales.

4. Likelihood ratio estimates are sometimes less extreme than odds estimates.

Other conclusions are also reviewed. Implications of these conclusions for the design of probabilistic information processing systems and for further research on response modes for information processing are discussed.

011313-6-T

Report Abstract 6

Relative Variance Preferences in a Choice-Among-Bets Paradigm

Raymond C. Seghers, Dennis G. Fryback, and Barbara C. Goodman

Abstract

The validity of the prime assumption of proper scoring rules (PSR), that people maximize subjectively expected value (SEV), was tested in the case where SEV was assumed to equal EV. A choice-among-bets paradigm was used in which the lists of bets conformed to the requirements of a PSR. Both real and hypothetical payoff conditions were used, and in addition, EV, variance, and odds of the gambles were systematically varied. Of the 12 Ss only 3 tended to maximize EV under both real and hypothetical payoff conditions, while relative variance preferences can account for the decision strategies of the other Ss. Inferred strategies were simpler and more consistent during the real payoff sessions. The effect of the gambles' properties was idiosyncratic and no overall conclusions were drawn. The use of the list of bets generated by a PSR as a response mode for inferring subjective probabilities is questioned because of the weakness of the SEV maximization assumption in this context.

011313-3-T

Report Abstract 7

BAYESIAN DATA ANALYSIS OF GAMBLING PREFERENCES

Dirk Wendt
University of Michigan

Abstract

This paper emphasizes the use of Bayesian data analysis for experiments with choices among gambles. In an introductory example, the method is illustrated by a comparison of two learning theories. Special problems arise with the analysis of data from decision making experiments which assume deterministic choice models which cannot be handled by Bayesian analyses. Several ways around these difficulties are suggested, discussed, and demonstrated on two sets of data from choice-among-gambles experiments.

Report Abstract 8

Some Criticisms of the General Models Used in Decision Making Experiments

Dirk Wendt

Abstract

The general normative model of expectation maximization is outlined and criticized for several reasons. It may not be appropriate as a normative model in a variety of situations where it is assumed to be rational. Some of its conditions, e. g., independence of evaluation-of-aspects and probability-revision cues, and correctness of the simple additive utility model, may not be met. Moreover, deterministic models may be too strong to predict human behavior properly. Perhaps they should be replaced by probabilistic ones. The emphasis of this paper, however, is not to doubt the applicability of the model in principle but rather to point at some problems where more research is needed.

011313-9-T

Report Abstract 9

TOWARD AN ERROR THEORY FOR PIP: INFERENCE BASED ON AN ALTERNATIVE FORMULATION OF THE DATA SPACE

Dennis G. Fryback and Ward Edwards

Abstract

Probabilistic Information Processing (PIP) systems, as currently conceived, use experts' intuitive judgments about the diagnostic impact of individual data as inputs for mechanical aggregation by Bayes's theorem. Past research has shown that the posterior odds output by PIP are much more extreme than those arrived at via human aggregation. Because of this superior efficiency PIP-type processing of fallible data has been recommended as an important tool for decision making. The present paper questions the uncritical use in PIP of estimated likelihood ratios as if they were veridical. A theory is developed which incorporates into the inferential process the inherent variability of human judgment. The resulting effect is a decrease in the posterior odds given by PIP. Employing specific distributional assumptions, a numerical example is given that shows the possible magnitude of this decrease. Application of the present results and their implications for further theoretical and empirical research are discussed.